Lieff Cabraser Heimann& Bernstein
Attorneys at Law

Susman Godfrey l.l.p.
a registered limited liability partnership

CDAS

September 27, 2024

VIA ELECTRONIC FILING

Hon. Ona T. Wang
Daniel Patrick Moynihan
United States Courthouse
500 Pearl Street
New York, NY 10007-1312

RE: ***Authors Guild et al. v. OpenAI Inc. et al.*, No. *1:23-cv-8292***
Alter et al. v. OpenAI Inc. et al.*, No. *1:23-cv-10211

Dear Magistrate Judge Wang:

On behalf of the parties in the above-referenced actions and pursuant to the Court's September 13, 2024 Order re Motions to Compel & Status Conference ("Order"), Dkt. No. 202, Counsel for the Author Class Plaintiffs hereby submit this **joint** status letter to apprise the Court of the progress of the resolutions of the issues in Dkt. Nos. 78, 168, and the status of the *Basbanes v. Microsoft Corp., et al.,* No. 1:24-cv-0084 matter, as directed in Dkt. No. 202.

A. TOPICS AT ISSUE IN THE ORDER

Dkt. No. 78: RFAs re: Inclusion of Class Works in Defendants' Training Datasets: On Friday, September 20, 2024, the parties met and conferred to discuss potential search strategies as directed by the Court. Plaintiffs are working on robust search proposals for OpenAI's consideration. The parties expect to continue their discussions in the coming weeks and in advance of the upcoming October 30, 2024 conference.

Plaintiffs' Position:

One issue that has hampered discussion relates to two books datasets (called "books1" and "books2") that OpenAI used to train GPT-3 and 3.5. These datasets are specifically identified as training datasets in OpenAI's paper on GPT-3 entitled "Language Models are Few-Shot Learners." *See* https://arxiv.org/pdf/2005.14165 at 8-9. Early in discovery, in response to Plaintiffs' request that these datasets be produced, OpenAI said "books1 and books2" were used to train GPT-3 and 3.5 but were "deleted in or around mid-2022 due to their nonuse." *See Authors Guild,* 23-cv-8292, Dkt. 143-4. OpenAI recently represented that it located a copy of two books datasets and produced those datasets, but it has been unable to confirm whether these correspond to the books1 and books2 datasets or if the datasets which it has produced were used to train GPT-3.5. To ensure that any searches run are complete and efficient, it is imperative that OpenAI confirm, as a first step, that it has produced and/or made available for inspection all datasets used to train its models, including books1 and books2. To the extent that books1 and books2 were irretrievably deleted—and that the newly produced "books" datasets do not correspond to books1 and books2—confirmation of that fact too will help the parties proceed with discovery efficiently.

OpenAI's statement that, at the September 20 meet and confer "Plaintiffs admitted that they had neither searched the data that OpenAI produced on September 16" (consisting of two large training datasets), nor developed any search proposals for OpenAI to consider" is false. On that meet-and-confer, Plaintiffs communicated their views on the search protocol based on its initial review of the produced datasets. Plaintiffs noted that their protocol would involve at least a two-step process of matching text to titles of the Class Works, then searching surrounding areas for text which matched excerpts from class works. Plaintiffs have reviewed the data OpenAI produced on September 16, made that clear to Open AI during the meet and confer, and continue to work with technical experts to propose a full search protocol which Plaintiffs expect to provide to OpenAI next week.

OpenAI's statement that Plaintiffs could have sent their experts to inspect the training data in a highly controlled environment in Defense counsels' offices at any point since "June" is also inaccurate. OpenAI's counsel required agreement on such a training data inspection "protocol" before Plaintiffs would be permitted to inspect the training data. *See* Dkt. 143-4 at 2 (noting that training data was "available for inspection subject to agreement on an appropriate inspection protocol. . . "). The parties submitted their joint stipulation regarding training data inspections on September 11, 2024. *See* Dkt. 200. Plaintiffs' technical experts are scheduled to begin inspecting the training data on-site in early October.

Lastly, OpenAI's suggestion that Plaintiffs have not served RFPs requesting the training data is perplexing. Plaintiffs served RFPs requesting exactly that on January 29, 2024. In OpenAI's February 28th responses, it confirmed that it "will make available for inspection . . . the pretraining data for models used for ChatGPT." Plaintiffs also requested, specifically, that OpenAI produce the books1 and books2 datasets, indeed it was that request that prompted OpenAI's counsel to state that these datasets had been deleted. *See* Dkt. 143-4 at 1-2. Now, seven months later, OpenAI refuses to confirm (A) whether it has produced books1 and books2 (or whether it destroyed them), and (B) thus whether it has either produced or made available for inspection all the relevant training data including books1 and books2.

OpenAI's Position:

At the September 12 status conference, the Court instructed Plaintiffs to explain what they want OpenAI to search to enable OpenAI to assess the Plaintiffs' RFAs. The book title? Author? First paragraph? Or a few scattered sentences? Although the parties have conferred, Plaintiffs still have not provided this information. If Plaintiffs actually believed it was "imperative that OpenAI confirm, as a first step, that it has produced and/or made available for inspection all datasets used to train its models," then Plaintiffs would have served requests for the production or inspection of those datasets. They did not. Instead, they sought to take *indirect* discovery of training data by serving the Requests for Admission that are the subject of Dkt. No. 78.[1] Nevertheless, despite Plaintiffs' inconsistent approach, OpenAI has made training data for the models-at-issue available since June of this year and continues to do so; Plaintiffs have not once taken the opportunity to review this data.

Two weeks ago, the Court ordered the parties to meet and confer so that Plaintiffs could identify what "we-want-them-to-search-for[s] or how [plaintiffs] want the search to progress." Hearing Tr. at 14. When the parties convened a week later, Plaintiffs admitted that they had neither searched the data that OpenAI produced on September 16 (the books datasets that OpenAI recently produced), nor developed any search proposals for OpenAI to consider.[2] (And again, neither have Plaintiffs looked at the datasets made available for inspection in June.) As of the date of this filing, OpenAI is still waiting for Plaintiffs to share search proposals. If Plaintiffs need more time, that is fine. But red herrings in the statement above do nothing to resolve ECF No. 78.

If Plaintiffs are looking for the datasets used to train GPT-3, they have either been produced or made available for inspection. Recently, Plaintiffs asked for clarification regarding a March 2024 letter, in which OpenAI relayed its *then*-current understanding based on its investigation to date. As OpenAI informed Plaintiffs, its investigation has continued to surface new documents and data that have been provided on an ongoing basis. Plaintiffs were told to rely on this information rather than outdated summaries from months ago. OpenAI also stated that it would consider Plaintiffs' specific question regarding the training datasets used for GPT-3.5, and it has been doing so. But Plaintiffs do not need this information to share their proposed search strategy for the RFAs at issue here.

Despite all the finger pointing in Plaintiffs' statement, the parties are in agreement on where they stand today: Plaintiffs still haven't told OpenAI how they propose to search the training data. OpenAI looks forward to receiving Plaintiffs' search proposals and will continue to work in good faith to resolve this issue without further burdening the Court.

<u>Dkt. No. 168: Dispute over Discovery from Microsoft Regarding "Training"</u>:

Plaintiffs' Position:

At the hearing before Your Honor on September 12, 2024, the Court directed Microsoft to "[m]eet and confer" to, among other things, "[c]onfirm, without the disclosure of work product, that the interviewed witnesses or witnesses in [Microsoft's] investigation [to search for training-

[1] Plaintiffs argue that they served RFPs seeking this in January. Not so. Instead, Plaintiffs served several overbroad RFPs in January; none of them specifically sought production of training data. *See, e.g.*, Plaintiffs' RFP No. 19 seeking "DOCUMENTS and COMMUNICATIONS reflecting or discussing how YOU accessed any commercial works of fiction or nonfiction, INCLUDING FICTION CLASS WORKS AND NONFICTION CLASS WORKS, used to train CHATGPT."

[2] Plaintiffs' suggestion to the contrary is inaccurate. Plaintiffs argue above that they had conducted an "initial review" prior to the parties' September 20 meet and confer. This is not a search.

related documents] are in the custodian list." Hr'g Tr. at 50:2-7. The Court also directed Plaintiffs to "talk to [Microsoft]" about conducting a custodial Rule 30(b)(6) deposition about the location and storage of responsive documents.

Consistent with the Court's order, the parties met and conferred on September 18 and Plaintiffs requested Microsoft's confirmation that the custodians interviewed or covered by Microsoft's investigation of training-related documents were included as document custodians in this case. Plaintiffs followed up by email on September 20. To date, Microsoft has not responded to this request, and has not confirmed the inclusion of the relevant custodians.

Plaintiffs have also requested custodial Rule 30(b)(6) depositions of both Microsoft and OpenAI, without prejudice to a later Rule 30(b)(6) deposition on the substantive issues in the case. The parties are addressing this issue in the context of the deposition coordination protocol.

On September 25, the day before the parties were set to exchange this letter, Microsoft supplemented its responses to a subset of the disputed RFPs. Many of these supplemental responses re-raise the disputed objection or raise new objections. The supplemental responses do not address RFP Nos. 22, 29, and 39, nor do they confirm whether the interviewed custodians' files are part of Microsoft's ESI collection. The parties have not yet conferred on these supplemental responses.

Microsoft's Position:

Dkt. 168 sought to compel further responses regarding a dozen document requests where Microsoft interposed an objection that, based upon its investigation, it did not train OpenAI's models and did not possess OpenAI's training datasets. After the hearing and a further meet and confer, Microsoft served supplemental responses clarifying its objections. The supplemental responses certify where Microsoft was refusing to undertake another search because it had already reasonably investigated the issue (three requests), and confirm where it was not limiting the scope of its search (nine requests).

Accordingly, Plaintiffs received the relief they requested in Dkt. 168 with the supplemental responses. As the Court noted at the status conference: (1) Microsoft agreed to search for and produce documents about training in response to many document requests; and (2) Plaintiffs will be able to review those documents and take depositions about that subject with fact witnesses. Hr'g Tr. at 50:7-11 ("I heard already that Microsoft is using 'train' or 'training.' If there's other terms or other ways you want to refine those searches, let that be an iterative process, with the idea that you will be doing depositions after you've gotten enough documents that make it worth doing the depositions.").

The service of supplemental responses should have resolved this issue without implicating Microsoft's work product. The Court recognized that the disclosure of the identity of witnesses in an investigation can implicate work production information. Hr'g Tr. at 46:22-47:13; 49:24-50:17. Particular witnesses in a prior investigation may or may not have information relevant to the question at hand, so the inclusion (or not) of those witnesses as custodians does not address the issue.

Nor have Plaintiffs made any showing that Microsoft's present investigation is inadequate. "[A] plaintiff is not entitled to conduct discovery that is solely relevant to the sufficiency of the adversary's document production without first identifying facts suggesting that the production is deficient." *Orillaneda v. French Culinary Inst.*, No. 07-cv-3206, 2011 U.S. Dist. LEXIS 105793, at *24 (S.D.N.Y. Sept. 19, 2011). When asked during meet and confer what basis they had for making such an assertion, Plaintiffs declined to offer any basis whatsoever.

Plaintiffs have identified no deficiency in the scope of Microsoft's production that would warrant the extraordinary remedy of discovery about discovery while discovery is still ongoing and months from conclusion. *See, e.g., Winfield v. City of New York*, No. 15-cv-05236, 2018 U.S. Dist. LEXIS 22996, at *12 (S.D.N.Y. Feb. 12, 2018)) ("When the discovery sought is collateral to the relevant issues (i.e., discovery on discovery), the party seeking the discovery must provide an 'adequate factual basis' to justify the discovery, and the Court must closely scrutinize the request 'in light of the danger of extending the already costly and time-consuming discovery process ad infinitum.'") (*quoting Mortg. Resol. Serv'ing, LLC v. JPMorgan Chase Bank, N.A.*, No. 15-cv-0293 2016 U.S. Dist. LEXIS 91570, at *20 (S.D.N.Y. July 14, 2016)); *Stephens Inc. v. Flexiti Fin. Inc.*, No. 18-cv-8185, 2019 U.S. Dist. LEXIS 127896, at *5-6 (S.D.N.Y. July 31, 2019) (same).

***Basbanes* Stipulation:**

On behalf of the Author Plaintiffs, Counsel for *Basbanes*, and Defendants, Author Plaintiffs submitted a Stipulation and [Proposed] Order on September 26, 2024. Dkt. No. 206.

B. PLAINTIFFS' POSITION ON ADDITIONAL TOPICS

"For-Profit Conversion," Licensing Negotiations, and OpenAI Models Subject to Discovery:

Plaintiffs' Position:

Counsel for the Author Plaintiffs joined the meet and confer on issues held between OpenAI and the New York Times / Daily News Counsel on September 20, 2024 regarding *NYT* Dkt. 128 (scope of models at issue), and *NYT* Dkt. 141 (formation of the OpenAI for-profit entity and licensing agreements/communications). Given the similarity of these issues to topics which the Author Plaintiffs and OpenAI have been discussing, Author Plaintiffs seek to streamline and coordinate the parties' discussions on these issues.

OpenAI's Position:

Inclusion of these issues in the parties' joint letter is improper. There is no pending motion on these issues, and the Court did not direct the parties to address them here. *See AG* ECF

No. 202. Nevertheless, OpenAI will continue to meet and confer with Plaintiffs on all discovery matters, including these ones.

Respectfully submitted,

LIEFF CABRASER
HEIMANN & BERNSTEIN,
LLP

/s/ Rachel Geman
Rachel Geman

SUSMAN GODFREY
LLP

/s/ *Alejandra Salinas*
Alejandra Salinas

COWAN DEBAETS
ABRAHAMS & SHEPPARD
LLP

/s/ Scott J. Sholder
Scott J. Sholder

MORRISON & FOERSTER,
LLP

/s/ Joseph C. Gratz
Joseph C. Gratz

LATHAM & WATKINS
LLP

/s/ *Elana Nightingale Dawson*
Elana Nightingale Dawson

KEKER, VAN NEST &
PETERS LLP

/s/ Thomas E. Gorman
Thomas E. Gorman

ORRICK HERRINGTON
& SUTCLIFFE, LLP

/s/Annette L. Hurst
Annette L. Hurst